

An Applied Evaluation of SNOMED CT as a Clinical Vocabulary for the Computerized Diagnosis and Problem List

Henry Wasserman, MS¹ and Jerome Wang, MD, FAAP^{1,2}

Enterprise Information Services¹

Departments of Medicine and Pediatrics²

Cedars-Sinai Health System

Los Angeles, California

The use of a standardized controlled terminology allows diverse systems and applications throughout the enterprise to translate data. In developing a customized enterprise-wide vocabulary for clinical terminology, we implemented SNOMED CT as a base vocabulary, while facilitating the addition of site-specific clinical terms or concepts not represented in SNOMED CT. In this paper, we evaluate the breadth of SNOMED CT terms and concepts for the coding of diagnosis and problem lists by clinicians within a computerized physician order entry (CPOE) system. Clinicians selected diagnosis and problem list terms from a lexicon based on SNOMED CT, submitting requests for clinical terms that were not found in the controlled vocabulary. For each "missing" term, we assigned one of four mapping types, representing the relationship of this new terminology entry to the SNOMED CT reference terminology. Our results show that the majority of diagnosis/problem list terms (88.4%) were found in SNOMED CT. Of the 145 missing terms, only 20 represented significant concepts missing from SNOMED CT, resulting in concept coverage of 98.5%. Our results show that SNOMED CT is a relatively complete standardized terminology on which to base a vocabulary for the clinical problem list.

INTRODUCTION

Integrating a controlled clinical terminology with the ability to clearly specify medical language within a computer-based patient record is an ongoing challenge. A standardized terminology supports efficient indexing and processing of patient data, and is an essential element for the implementation of knowledge-based clinical decision-support, data retrieval and aggregation. This is accomplished by exploiting pre-defined semantic relationships, both hierarchical and non-hierarchical in nature. In order to reap the benefits of such a clinical vocabulary, a number of research and academic centers have developed customized terminologies over the course

of many years. However, since significant expertise and resources are required to create and maintain a customized clinical terminology, this approach may be impractical for other institutions that choose to customize an industry-supported and maintained standardized vocabulary.

Investigators have performed controlled evaluations of the coverage of standardized clinical vocabularies in a variety of clinical domains. These studies included a comparison of ICD-9-CM, ICD-10, SNOMED III, Read, UMLS, CPT and NANDA.^{1,2} However, with the merging of SNOMED RT and the Read codes into the first release of SNOMED CT in early 2002, a centrally standardized and maintained clinical terminology has become commercially available.³ The purpose of this study is to evaluate SNOMED CT in its coverage of terms and concepts needed for the comprehensive encoding of a clinical problem and diagnosis list in a real-world clinical setting.

METHODS

Cedars-Sinai Medical Center (CSMC) is an 820-bed non-profit tertiary hospital at which a CPOE system was recently implemented in October 2002. The data collected in this study represent entries into the diagnosis/problem list for each patient admitted into the hospital during a four-month period.

In the course of this study, we felt it was important to distinguish between a "display term" and the concept underlying the display term. For the remainder of the paper, we follow the vocabulary of Rocha et. al. in calling these display terms *surface forms*, since the caregiver sees only the concept's "surface" representation (the display term), while the abstraction of the concept is hidden within the term.⁴

Pre-CPOE Implementation Lexicon Creation

Since we wished to retain diagnosis terms frequently in use at CSMC, we required a lexicon that supported

both traditional diagnosis terminology as well as a large variety of “standard” CSMC diagnosis terms. SNOMED CT, First Edition (January 2002)³ was chosen as the base vocabulary, for both its large lexicon of clinical terminology and well-defined hierarchical structure. We chose the simple strategy of creating a base lexicon of all concepts and synonyms under the “disease” concept node (58,807 concepts, 23,800 synonyms) in the SNOMED CT hierarchy, with several additions. To support traditional diagnosis terminology, we mapped 148 diagnosis terms frequently used at CSMC to their corresponding concept id. Thus, we added a set of CSMC-specific synonyms by mapping a new display term to an existing SNOMED CT concept.

Selection of Diagnosis and Problem List Terms within the CPOE

The data for this evaluation represents a compilation of the terms entered by caregivers at CSMC as either a *diagnosis* or *problem* in the CPOE. If a physician wishes to write orders for an inpatient, the CSMC clinical information system requires that caregivers specify a primary diagnosis, and stores the diagnosis in a local database. The caregiver also has the option of specifying multiple secondary diagnoses or “problems” for each patient, although these are not required by the system.

The CSMC clinical information system allows the caregiver to select a diagnosis or problem in two ways. The first selection method, shown in Figure 1, allows the caregiver to choose a diagnosis from a pull-down menu: the caregiver first chooses her clinical specialty, and then selects the diagnosis from a list of frequently chosen diagnosis terms. The second selection method, shown in Figure 2, allows the caregiver to enter a free-text query of a subset of the SNOMED CT database. The system returns all “hits” for the given free-text query, using a search algorithm implemented in SQL to generate potential matches. The user can then select any one of the returned terms by clicking on it, or can view the term’s “parents” in SNOMED CT, by clicking the less specific link next to the desired term. Likewise, clicking the more specific link brings up a list of a term’s “children.” These hierarchical relationships are well described in the SNOMED CT database, and can be easily accessed via an SQL query. Using one of these two selection methods, the caregiver’s selected term populated a local database that held patient information.

The screenshot shows a web interface for selecting a diagnosis. At the top, there is a 'Search' button and a 'Diagnoses' tab. Below this is a section titled 'Common Diagnoses'. A 'Clinical Specialty' dropdown menu is set to 'Pediatrics'. Below that is a 'Diagnosis' dropdown menu. To the left of the dropdown, there are links: 'Find a problem or', 'Name', 'You can choose a specific to find a', 'Search Results', and 'Click to add this'. The dropdown menu is open, showing a list of diagnoses: 'Acute appendicitis', 'Altered mental status', 'Apnea', 'Apparent Life-threatening event (ALTE)', 'Aspiration pneumonitis', 'Asthma' (highlighted), 'Behavioral and emotional disorder with onset in childhood', 'Bronchiolitis', 'Cellulitis', 'Congestive heart failure', and 'Croup syndrome'.

Figure 1: Selection of diagnosis using a pull down menu of frequent diagnoses.

The screenshot shows a web interface for searching for a diagnosis. At the top, there is a 'Find a problem or diagnosis' section. Below this is a 'Name' input field with the text 'ac asthma' and a 'Go' button. Below the input field, there is a message: 'You can choose a search result below if it is highlighted. If the result you want is not highlighted, click on the link "Less specific" to find a related result.' Below this is a 'Search Results' section. It contains a table with three columns: 'Click to add this problem or diagnosis', 'Less specific', and 'More specific'. The table lists several asthma-related terms, each with links to 'Less specific' and 'More specific' results.

Click to add this problem or diagnosis	Less specific	More specific
Accidental poisoning by herbal asthma mixture	Less specific	More specific
Acute asthma	Less specific	More specific
Acute exacerbation of asthma	Less specific	More specific
Acute severe asthma	Less specific	More specific
Asthma attack	Less specific	More specific
Cardiac asthma	Less specific	More specific
Exacerbation of asthma	Less specific	More specific
Extrinsic asthma with asthma attack	Less specific	More specific
Factitious asthma	Less specific	More specific
Intrinsic asthma with asthma attack	Less specific	More specific

Figure 2: Selection of diagnosis using free-text query and SNOMED CT navigation.

Addition of Missing Terms

In order to support hospital-specific or other terminology not described by the SNOMED CT database, as well as the changing needs of the clinician⁵, we encouraged caregivers who could not find a desired diagnosis term to submit the new term similar to a process first described by Warren et al.⁵ Options existed to immediately contact our on-site 24-hour “command center” in person, by email, or by phone. The terms were immediately forwarded to one of the authors (JKW), and a rigorous search within the SNOMED CT base vocabulary as well as our customized terminology was performed to determine if the requested term was truly “absent” from SNOMED CT, or if the searcher failed to find the appropriate existing term. We defined a term as absent if it offered a concise description for a diagnosis or problem, and differed significantly from all existing SNOMED terms. If a term was deemed

“absent,” we then placed it into one of four categories⁶:

1. Synonym – CSMC
2. New Leaf
3. New Leaf with multiple stems
4. Graft to branch

These four categories were adapted as an alternative to more general categories of measurement, such as “Exact meaning,” “Related concept,” and “No related concept” presented in previous studies.⁷ We felt that the above categories gave an accurate and comprehensive measurement of the significance of the absence of a term, in addition to making the process of adding terms to the terminology more efficient. The categories were chosen such that they cover all possible ways that a new concept node might be inserted into the SNOMED CT hierarchy. We describe each category in detail, facilitating a deeper discussion of the coverage of SNOMED CT in the Results and Discussion sections of the paper.

Synonym—CSMC: This category represents missing surface forms: terms that do not exist in the SNOMED CT database, but have a direct mapping to a SNOMED CT concept. For example, we added the term “abnormal pap smear,” which can be directly mapped to the SNOMED term “abnormal cervical smear.” While SNOMED CT contains the appropriate concept, it does not contain the desired description term, so we simply add the new surface form (abnormal pap smear) to the list of diagnoses, and map this term to an existing SNOMED CT concept (abnormal cervical smear).

New Leaf: While an absent synonym can be remedied by simply adding a surface form, a missing concept represents a more significant absence. A new leaf term requires the generation of an entirely new concept, along with the surface forms and the structure and relationships associated with that concept (relationships, descriptions, cross mappings, etc.). Terms in this category cannot be mapped to an existing SNOMED concept, and therefore require generation of a new concept node in SNOMED CT. An example of a *New Leaf* term found in this study is “iatrogenic pneumothorax,” which is a child of the term “Pneumothorax.”

New Leaf with multiple stems: A term in this category is identical to a *New Leaf* term, with the exception that the term has multiple parents, or “stems” attached to the SNOMED CT tree. For example, the surface form “nosocomial pneumonia” (closest match was “nosocomial infectious disease”) does not exist

in SNOMED CT, and should have multiple parents, since it represents a “child” of two concepts—it belongs in the hierarchy as a child of both “bacterial pneumonia” and “nosocomial infectious disease.” Terms in this category require more work than *New Leaf* terms, since they must be added to multiple places in the SNOMED CT hierarchy.

Graft to Branch: Terms in this category represent a gap in the SNOMED CT hierarchy: the term is absent, and the missing term has a parent relationship to one or more concepts. Thus, in order to add the concept to the hierarchy, we must “graft” it to some branch of the SNOMED CT tree. For example, a term in this category is “Sepsis from indwelling medical device”, which would be a parent of the SNOMED CT concept “Tracheostomy sepsis, as well as a missing concept “Line sepsis”. A concept corresponding to “Sepsis from indwelling medical device” does not currently exist in SNOMED CT, but a parent of this concept (“Systemic infection”) does. This concept fits between existing parent and child concepts, and must be “grafted” to an existing branch of the SNOMED CT hierarchy in order for the term “Line sepsis” to be modeled correctly.

Term Analysis

The final step in the data processing cycle involved the analysis of all selected diagnosis terms through a variety of software tools. We developed these tools in the Java programming language, using an object-oriented design structure which matched the structure of the SNOMED CT concept entity. After extracting the diagnoses from the patient database, we used the surface form of each diagnosis as an entry point into the SNOMED CT description table. This table allowed us to determine the concept identifier, description type, and description status of each term.⁸

RESULTS

Over a four-month period (10/24/02-1/23/03), physicians encoded a total of 8,378 diagnoses and problems. Table 1 shows that 1,266 of these diagnoses had unique surface forms, representing a total of 1,105 distinct concepts. 56% of diagnoses and problems were selected through a pick list, and the 10 most frequent diagnoses accounted for 40.5% of all diagnoses.

Of all 3,663 diagnoses selected by free-text search, 826 contained abbreviations. As expected, this result shows that clinicians frequently search for long diagnosis names by typing their common abbreviations, such as “BPH” for diagnosis “benign

prostatic hypertrophy.” Of the 528 unique synonyms, 169 contained an abbreviation. The most recent version of SNOMED CT contains a large number of abbreviations as synonyms, and also provides a tool for mapping abbreviations to concepts in the form of a “word equivalents” table.⁸

Table 1: Summary of collected diagnoses terms

	Total
Total diagnoses	8378
Unique surface forms	1266
Distinct concepts	1105
Selected by free-text search	3663
Selected from pick list	4715

Figure 3 shows a breakdown of unique surface forms by type, as chosen by clinicians as a diagnosis or problem in the CPOE. A relatively large number of synonyms were chosen, considering that SNOMED CT contains 2.4 times more preferred terms than synonyms (the July 2002 release contains 351,482 preferred terms and 145,010 synonyms). The large number of synonyms chosen shows the value of utilizing a wide variety of surface forms to match the variety of clinical language used to describe diagnoses and problems.

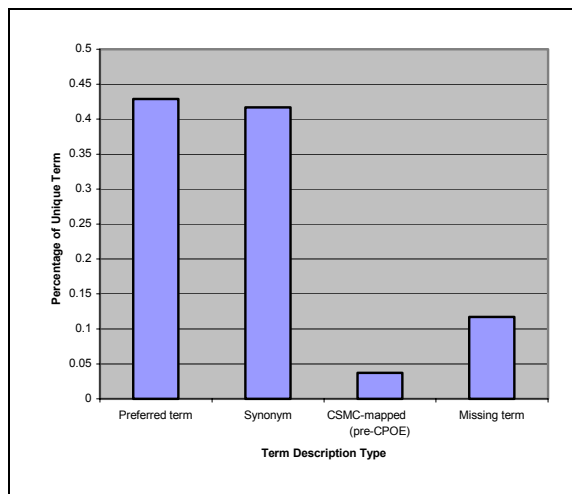


Figure 3: Breakdown of unique clinician-selected surface forms by type

Of all unique surface forms chosen, 11.6% represent terms added by the authors, loosely labeled as “missing terms.” However, it is important to note that the addition of surface forms of the “synonym” type was an extremely simple process. As a consequence of this simplicity, the authors sometimes added synonyms even when a sufficient surface form

already existed. For example, we added the synonym “AV block – 1st degree” for the concept “First degree atrioventricular block,” allowing the clinician to find the term using the abbreviation “AV”. Since the addition of synonyms could be rapidly completed, and the searcher could find the desired concept with slightly more time spent searching, we found these “misses” to be relatively insignificant, and therefore did not represent a true lack of coverage in SNOMED CT.

Table 3: Breakdown of missing terms by type

	Number of terms added
Synonym – CSMC	125
New Leaf	12
New Leaf with Multiple Stems	4
Graft to Branch	4

Table 3 shows that the “synonym” type dominated the missing surface forms. The remaining 20 terms represent more significant misses, requiring the addition of concepts and relationships to the SNOMED CT hierarchy. These represent concepts truly absent from SNOMED CT: diagnosis/problem terms for which no adequate description currently exists in the database. Only 1.5% of the unique surface forms coded by clinicians to describe a diagnosis/problem represented concepts that could not be found in SNOMED CT, resulting in a concept coverage of 98.5%.

In an examination of the missing terms, we found that nearly all of the *new leaf* terms were more detailed descriptions of an existing SNOMED CT concept. For example, the term “nursing home acquired pneumonia” is a more descriptive term for the SNOMED CT concept, “bacterial pneumonia.” 3 of the 4 *new leaf with multiple stems* could be expressed by combinations of 2 existing concepts. For example, the CSMC surface form “nosocomial pneumonia” could be represented by the combination of two SNOMED CT codes representing the individual terms “pneumonia” and “nosocomial infectious disease.” However, forcing the searcher to select multiple terms to describe a common problem forces the searcher to spend more time navigating the search interface.

DISCUSSION

The need for a patient-centric problem list is clear. Agencies such as the Joint Commission on Accreditation of Hospital Organizations (JCAHO) have mandated that hospitals maintain a longitudinal patient-centric problem list, while a 1991 publication

by the Institute of Medicine (IOM) states that the problem list is an essential component of the medical record, by improving the quality and coordination of patient care.⁹ Although paper-based problem lists can satisfy these requirements, the benefits of an computerized and encoded problem list entered by physicians at the point-of-care offers a tremendous potential for improved quality of care. A computerized problem list is often more readily accessible than the paper chart, and codified terms create an opportunity to implement clinical decision-support features, such as knowledge retrieval, error trapping, and clinical guidelines. However, attempts to implement such a coded system only highlight the challenges of creating and/or maintaining an underlying vocabulary with the breadth and depth of content within a computerized information system.

Therefore, a number of investigators have evaluated the coverage of a number of standardized clinical terminologies.^{1,2} These studies analyzed large sets of clinical terms submitted in a controlled research environment, seeking to analyze the completeness of clinical vocabularies in a variety of domains (such as diagnoses, findings, and procedures). Using similar methods, others have evaluated the UMLS and ICD-9 code sets specifically in the context of the problem list.^{10,11} However, it was evident that these vocabularies significantly lacked completeness in term coverage. To our knowledge, our study is the first to evaluate the coverage of SNOMED CT for purpose of coding terms for the medical problem list. We have purposefully conducted this study in the “real-world” setting by testing SNOMED CT content coverage against a gold-standard, the practicing physician at the point-of-care.

Given the methodology, however, our study has a number of limitations. It is possible that clinicians coding medical diagnosis entries may have “settled” for terms that did not reflect exact matches to what was intended. Although we were unable to evaluate this occurrence in our study, it is important to note that problem list entries in the clinical record often do not reflect the most granular concept within a given hierarchy. It is also possible that a limited number of rare problem or disease entities were not encountered during the study setting, thereby adding a number of “missing terms”. However, we believe that these hypothetically missing terms would be limited given the large number of terms and hospital admissions sampled during the study, and would not impact the importance of our findings.

Our results suggest that SNOMED CT can be directly implemented as a base vocabulary for a clinician-coded patient-problem list within our medical center. The coverage of SNOMED CT terms and concepts in this context is above 90%, and the limited resources required to support just-in-time entry of missing terms and concepts into our problem list lexicon is further evidence of this.

REFERENCES

1. Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The content coverage of clinical classifications. *JAMIA* 1996;3(3):224-33.
2. Campbell JR, Carpenter P, Sneiderman C, Cohn S, Chute DG, Warren J. Phase II evaluation of clinical coding schemes: Completeness, taxonomy, mapping, definitions, and clarity. *JAMIA* 1997; 4(3):238-51
3. Spackman KA et al, eds. SNOMED Clinical Terms First Release. College of American Pathologists, Northfield, IL, 2002.
4. Rocha RA, Huff SM, Haug PJ, Warner HR. Designing a Controlled Medical Vocabulary Server: The Voser Project. *Computers and Biomedical Research*, 1994; 27(6):472-507.
5. Warren JJ, Collins JC, Sorrentino C, Campbell JR. Just-in-Time Coding of the Problem List in a Clinical Environment. *Proc AMIA Symp* 1998;:280-284
6. Campbell JR. Integrating Terminology Standards (SNOMED RT) Into a Clinical Information Environment. Scottsdale Institute Presentation, July 25, 2002.
7. Humphreys BL, McCray AT, Cheh ML. Evaluating the coverage of controlled health data terminologies: Report on the results of the NLM/AHCPR Large Scale Vocabulary Test. *JAMIA* 1997;4(6):484-500.
8. SNOMED Clinical Terms Technical Specification. College of American Pathologists, Northfield, IL, 2001. Available from: <http://www.SNOMED.org>. Accessed July 2002.
9. Dick, RS, Steen EB, eds. *The Computer-Based Record: An Essential Technology for Health Care*. 1991; Institute of Medicine, National Academy of Sciences.
10. Payne TH, Martin DR. How useful is the UMLS metathesaurus in developing a controlled vocabulary for an automated problem list? *Proc Ann Symp Comput App Med Care*. 1993;705-9
11. Payne TH, Murphy GR, Salazar AA. How well does ICD9 represent phrases used in the medical record problem list? *Proc Ann Symp Comput App Med Care*. 1992;205-9.